DOCUMENT RESUME

ED 215 575 FL 012 873

AUTHOR Whalley, Peter
TITLE A Partial Index of Text Complexity Involving the Lexical Analysis of Rhetorical Collectives. IET Text-Processing Paper 4.
INSTITUTION Open Univ., Walton, Bletchley, Bucks (England). Inst. of Educational Technology.
PUB DATE 80
NOTE 15p.; Paper presented at the International ALLC Symposium "Computers in Literacy and Linguistic Research" (6th, Cambridge, England, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cohesion (Written Composition); *Discourse Analysis; *Paragraph Composition; Statistical Analysis; Writing Evaluation
IDENTIFIERS *Text Structure

ABSTRACT

An alternative to the traditional surface measures of text complexity is put forward. The background to the study is an attempt to provide a useful analysis of educational courses being prepared on a text-processor system. The analysis is primarily concerned with one aspect of cohesion in text, the use of rhetorical connectives or items of conjunction as identified by text linguists. The aim is to tap some portion of the intended rhetorical structure as it is likely to be perceived by the reader, who must rely on the explicit signalling cues available. The analysis is made within the framework of a "staging" model of text production, which assumes a process of ordering and structuring of important notions. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Textual Communication Research Group

Institute of Educational Technology
The Open University
Walton Hall
Milton Keynes MK7 6AA

Milton Keynes (0908) 74066

IET

ED215575

IET Text-processing paper 4: A partial index of text complexity involving the lexical analysis of rhetorical connectives

Peter Whalley

Abstract An alternative to the traditional surface measures of text complexity is put forward. The background to the study is an attempt to provide a useful analysis of educational courses being prepared on a text-processor system. The analysis is primarily concerned with one aspect of cohesion in text, the use of *rhetorical connectives* or *items of conjunction* as identified by text linguists. The aim is to tap some portion of the author's intended rhetorical structure *as it is likely to be perceived by the reader*, who must rely on the explicit signalling cues available. The analysis is made within the framework of a 'staging' model of text production, which assumes a process of ordering and structuring of important notions.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Peter Whalley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL 012873

Presented at the Sixth International ALLC Symposium 'Computers in Literary and Linguistic Research', University of Cambridge, 1980.

Acknowledgement

I would like to thank Shirley Owen for her efforts to simulate all the programs that didn't work, and Maurice Dow for help with one that did.

Introduction

The context for this study is an attempt to construct an advice system for use in the production of low resource educational texts (Lefrere, 1980). Part of the assistance that such a system might usefully proffer would be a comparative analysis of textual complexity, or some measure of the development or emphasis given to a particular topic. In their detailed narrative studies, text linguists have developed analytic techniques for evaluating many of the factors involved in textual complexity. This study was intended to examine the possibility of mimicking such detailed analyses with the simplest analytic techniques of linguistic computing, and to evaluate its potential as a source of useful comment.

The Linguistic analysis of role relations in text

Text linguistic models of discourse suggest that prose may be characterised in terms of three basic dimensions 'content', 'cohesion' and 'staging' (Grimes' (1976) term, Halliday (1976) refers to 'topicalisation'). Content is simply the semantic component of text and is not an issue here. The basic notion of cohesion is the manner in which linguistic devices are used to bind segments of a discourse together. It is seen as being cumulative and linear rather than hierarchical, and is more concerned with introducing new information and keeping track of old, than with the actual content of the information .

Halliday discusses cohesion under the headings of reference, substitution, ellipsis, conjunction and lexical cohesion. Conjunction (a concern of this study) is considered to be different from the other cohesive relations. It is described as a "non-structural, text-forming relation" which "is encoded not in the form of grammatical structures but in the looser more pliable forms of linkages between the components of a text".

Halliday anticipates the automatic analysis of texts by computer, using cohesive reference schemes; emphasizing that the analysis should be a means to an end (in this case explication of the author's intended rhetorical structure). Questions that are thought likely to arise include: favouring of one type of cohesion by particular authors; systematic variation in the density of cohesive ties; and the

relation between cohesion and paragraph division.

Staging relations are concerned with expressing the speakers' perspective on what is being said. Grimes' notion of staging involves a hierarchical development of topics, height in the structure indicating the degree of prominence given to particular parts of a message. Several studies have shown the psychological validity of this distinction in terms of memory load and structural perception of texts (Meyer,1974 ;Clements,1976).

Discourse Production

Clements has developed a discourse production model based on the role of staging in the writing and reading processes. Topics are chosen and organised (staged well or badly) by the author, and Clements' experiments also indicate that staging plays an important part in any reception model. Until the reader has a complete command of the subject matter, the explicit signalling cues of the author's staging structure dominate any judgements concerning the relative importance of topics. It is for this reason that a 'surface' analysis of text structure need not be theoretically superficial.

Generally, the more explicit the relationship between new ideas and the previous text is made, the easier it is for the reader to follow. Bruce (1978) claims that a sign of good writing is to have a large store of connective operators that can be used to indicate precisely where each new idea fits into the discourse structure. (However he does point out that polished but empty prose results from the use of the connectives to cover up flaws in the content.)

An essential component of the staging dimension is the use of the coordinating and subordinating connectives. Hence the rationale for the present study; a statistical analysis of the density of rhetorical connectives between different parts of a text or texts. The connectives used in this analysis (Appendix A) are based on a table compiled by Clements, supplemented by others taken from Van Dijk and Halliday. Following Grimes (and also van Dijk), Clements groups the connectives into coordinators and subordinators, the essential distinction for Clements' study (and this one) being that the subordinators operate exclusively within-sentence.

There are several obvious limitations to this list. It does not include the many connective phrases ('due to', 'as a result',

'such...that', etc), although this can be simply remedied. Also without some fairly simple parsing capability, to detect lists and use as phrasal connective, the coordinator 'and' must be excluded. (Similarly simple parsing might pick up the use of the comma as contrastive connection.)

A more fundamental problem is that the coordinator-subordinator distinction is imperfect and text linguists do not agree as to the function of all of the connectives. Any purely statistical analysis will have problems when the expressions enumerated may validly take on different roles. However, for simple comparative studies it is sufficient to acknowledge these difficulties. Indeed if processing limitations were important, the list could be shortened and 'tuned'. If a more detailed parsing system were envisaged, Van Dijk (1977) has now provided a 'formal grammar of the natural connectives'. Previously the grammar of connectives had remained purely "ad hoc" ,as most text grammatical research has been directed toward intersentential relations (co-reference, pronouns etc).

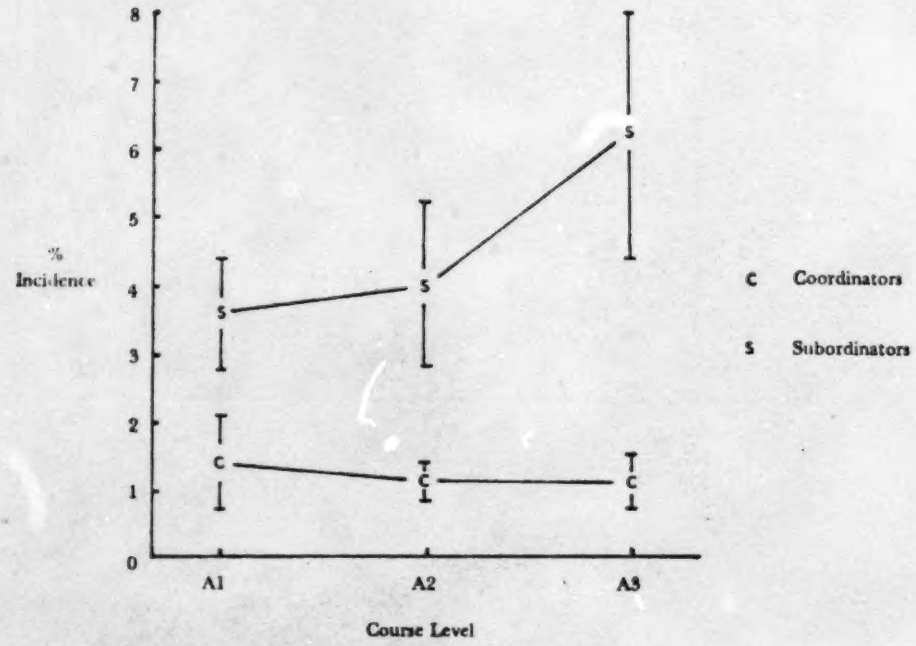


Fig. 1. Incidence of connectives across three levels of Arts courses.

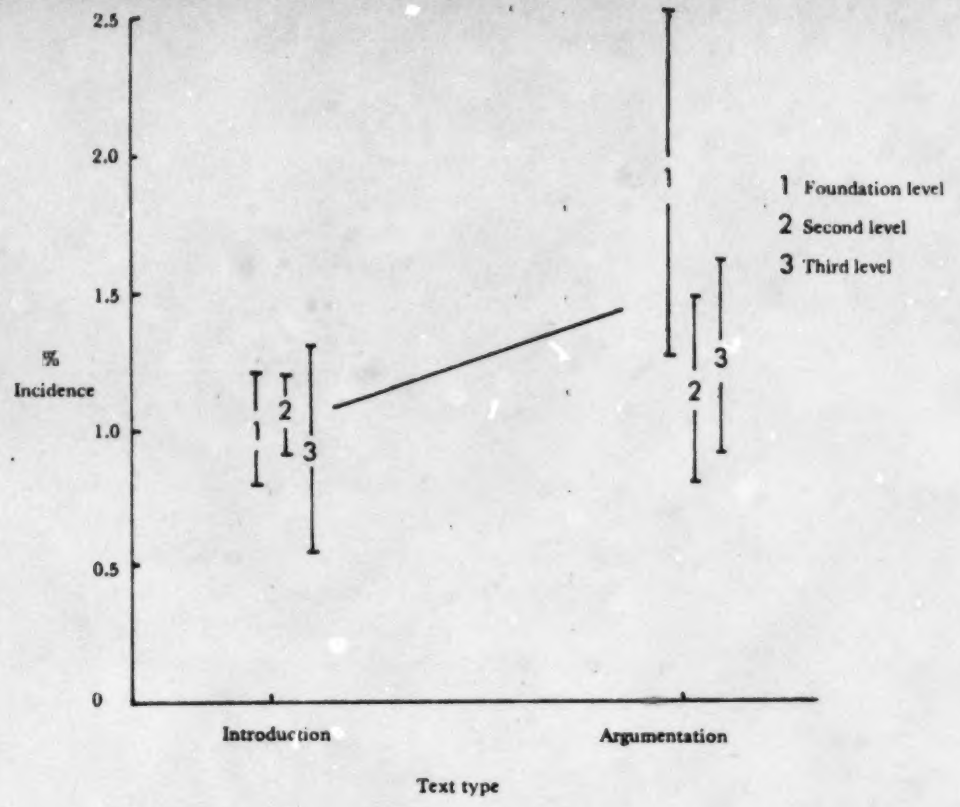


Fig. 2. Incidence of coordinators in various parts of a discourse (Arts courses).

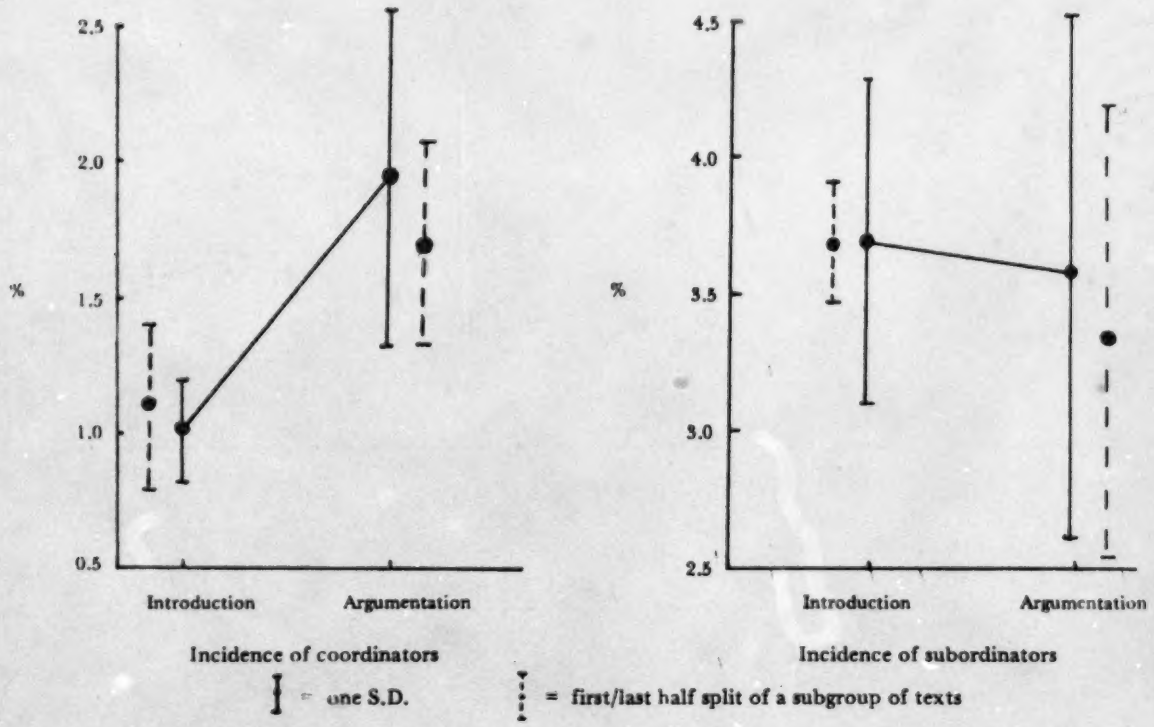


Fig. 3 Incidence of connectives in various parts of a discourse
(Arts course, foundation level)

Despite these limitations, the results of a simple statistical density count of the connectives across a range of texts indicates that the coordinator subordinator distinction is capable of providing useful comparative data. The analysis of 33 samples (each of about 1000 words) taken from three levels of Open University course is shown in Figures 1 and 2. It appears that the (within-sentence) subordinators may provide an index of syntactic complexity, and that the (between-sentence) coordinators can indicate the general form of discourse within a text. The second and third level samples were less homogeneous and obviously less well suited to a binary (introduction-argumentation) classification; this explains the greater variance. A more detailed analysis is being made using a greater number of text descriptors, on portions of the texts. A beginning-end division of the Foundation level texts resulted in a similar differentiation in coordinator density, figure 3.

Rhetorical connectives and paragraph structure

Halliday (1976) describes the ways in which a writer may use cohesion to signal texture, and how the reader will react to cohesion in his interpretation of texture and use it as a criterion for the recognition of text boundaries. Discontinuities or transitions in the text will be shown up by, "isolated sentences or other structural units which do not cohere with those around them even though they form part of a connected passage". Some writing is found to have rhythms of tight and loose texture, and Halliday considers the paragraph to be a device introduced to suggest just this kind of periodicity. In much written English a greater degree of cohesion is found within a paragraph than between paragraphs, however for some writing and authors it is found to be contrapuntal.

Werlich (1976) defines the concept 'text unit' as "those constituents of texts which enable the encoder to produce coherent and completed linguistic utterances of increasing extension beyond the sentence". Each text unit is marked by its own structure of coherence and completion and Werlich makes the prescriptive statement that there are five text types: book, part, chapter, section, paragraph, sentence. However, Grimes (1975) claims that "there is probably no means of establishing a limit on how many intermediate levels of organisation there can be between the paragraph and the discourse". Instances are a line of argument taking up several paragraphs and the idea of an

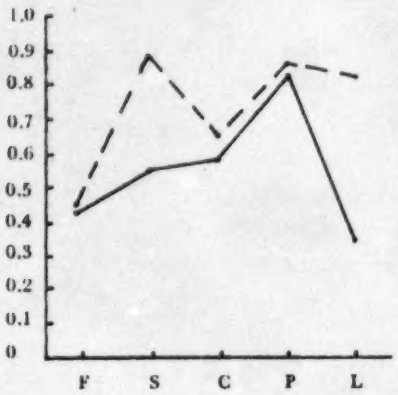
'episode' of several paragraphs being related to a "significant change of participants" ('issues' or 'concepts' could perhaps be substituted in non-dramatical works). Grimes gives some emphasis to the idea of 'planes of an overlay' as another kind of complexity intermediate between the paragraph and the discourse. The overlay introduces new information, at the same time highlighting common information between new and old planes.

Gindin (1978) cites various Soviet studies (both stylistic and statistical) concerned with the various intermediate continuous structural components of a text, the paragraph and the suprasentential entity (SE). Paragraphing and other graphic divisions are noted to have first been used as printer's devices and only later "acquired a motivation". The distribution of paragraphs can be thought of as "being more or less motivated", a 'good' division into paragraphs is considered to be one that is in accord with the inner structure of the text. The SE is viewed as a combination of sentences having a common 'microtheme' setting it off from the adjoining SE's and is likened to the long pause found in oral speech. Gindin claims that analysis of descriptive discourse shows that the initial nouns in all the sentences within a paragraph, except the first one in the SE, denote objects already mentioned.

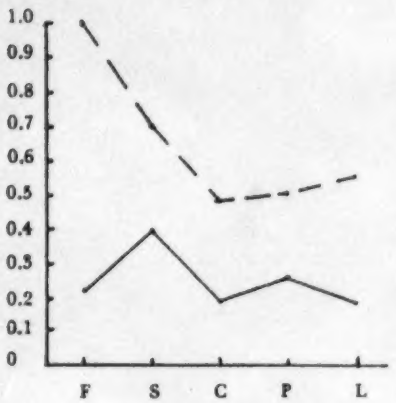
A statistical analysis of the paragraph distribution of connectives for several texts is presented in Fig 4. (This data has been adjusted both for varying numbers of sentences per paragraph and across texts). Even at a purely statistical level, Halliday's predictions appear to be borne out in the differences between authors and types of text. Some support for the existence of SE's is found in the co-occurrence of sub-titles (in 'New Scientist' articles) with paragraphs that do not conform to the general trend. It seems possible that the editor inserting sub-headings requires only to skim the text for cues such as those outlined by Gindin.

In a recent article, Van Dijk (1979) posits a tentative list of 'relevance assignment' signals ranging from typographic (e.g. sub-headings) and lexical devices to the rhetorical operations. (Despite reference to these signals as 'levels', it is probably best to conceive of them as a continuum as several of them act at 'all other levels'.) Any serious attempt to develop an index of rhetorical complexity would have to attempt to encompass as wide a range of these devices as possible. Extra-linguistic signalling has usually to be made explicit in text-processor systems (e.g. numbering systems, emboldening, footnotes etc) and is thus available for analysis.

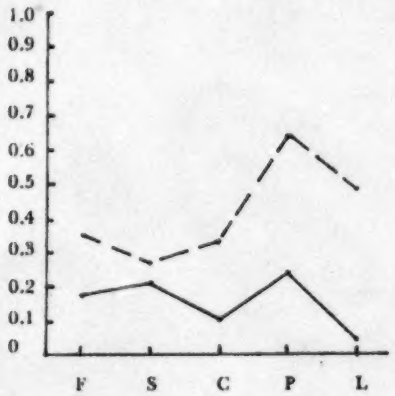
'New Scientist' article about the concept of species



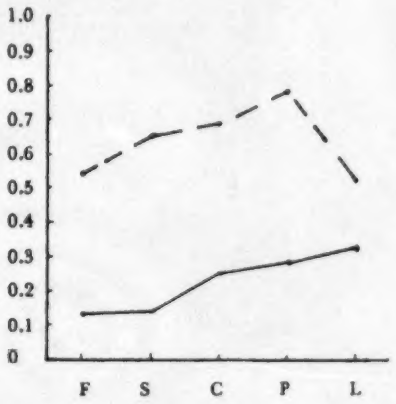
'Hi Fi News' article about the theory of amplifier-design



'Foundation' level course concerning evolution



Specialized journal article concerning species



'New Scientist' article concerning the mathematical concept e

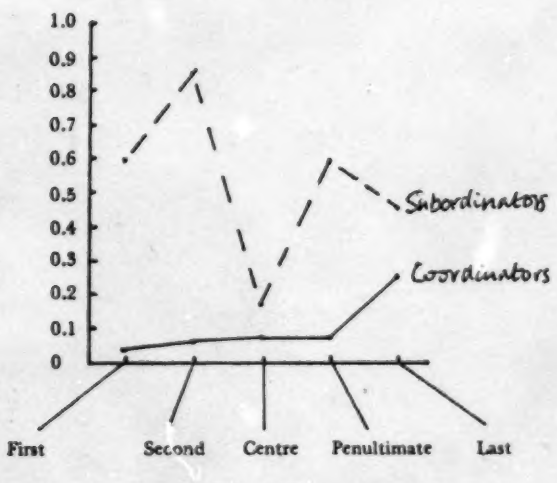


Fig. 4 Paragraph distribution of the number of connectives per sentence for various texts.

Despite a complete 'grammar of design' not being currently available (but cf Waller, 1980) a practical analysis might start from the functional equivalence of various typographical, lexical, and rhetorical systems when dealing with common problems such as side issues. Given the existence of text grammars for the linguistic cohesive devices it would seem possible to relate an analysis of the typographic cues to these. Indeed, contracts for American military technical manuals now specify simple design criteria along with the conventional readability guidelines.

Lexical analysis of connectives as an indicator of discourse topic

Given the results outlined, it was thought worthwhile extending the density of connectives 'analysis' to a study of the development of important topics in texts. This is effectively a selective word frequency analysis of a concordance output, and represents another attempt to see whether simple statistical analysis can provide useful information; or whether complex parsing of the cohesive references would be necessary.

The study was particularly concerned with educational texts, analysis of two complete Open University Foundation level Science units being made (each of 80-90 thousand words). The context of the key topic 'species', revealed no difference in the pattern of subordinators. However in the unit where species was defined in the concept glossary as 'being developed' ,the density of coordinators was significantly greater than that in the unit where species was made use of equally frequently (97 times versus 83), but where it was defined as having been introduced previously. This is shown in figure 5, as is the similar pattern for other less frequently occurring concepts (They each occur about ten times in the unit). Figure 6 indicates the effect on the results of increasing the length of context. It appears that the high density of connectives 'spreads' further around key concepts.

Additional concerns not developed here would include the change in development of key concepts between drafts or versions of the same course, and also the occurrence of 'mis-staging'. This might occur where the author places undue emphasis on som problem in the exposition in deference to academic peers, leaving the students open to confusion as to the important issues for them.

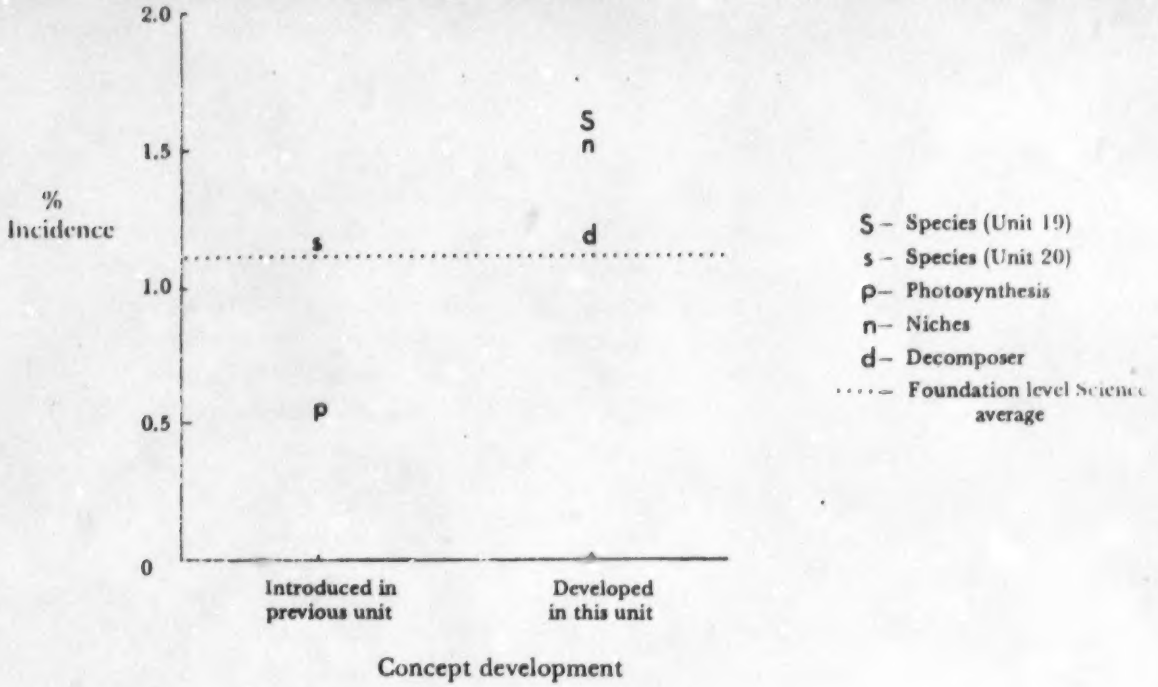


Fig. 5 Connectives as an indicator of topic development (coordinators)

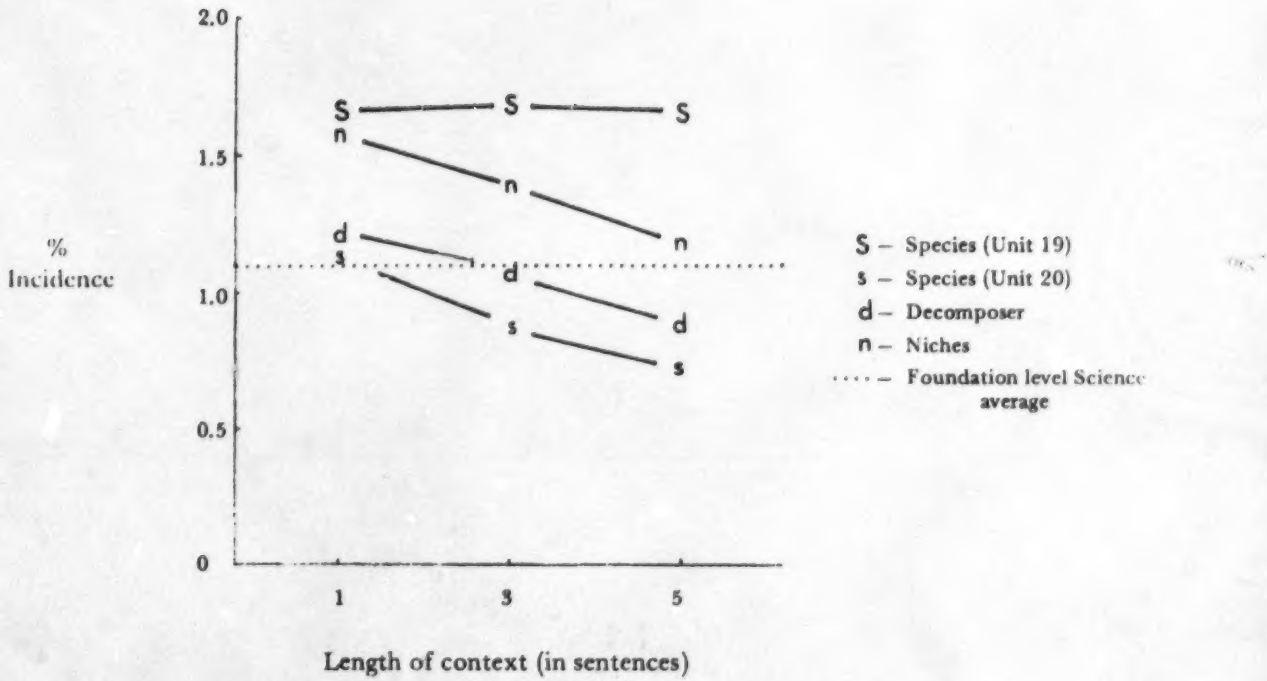


Fig. 6 The effect of context length on the density of coordinators

Conclusion

The results presented here show that a lexical analysis of the coordinating and subordinating connectives used in a text provides some indication of text complexity, and thus potential comprehension difficulty. A similar simple analysis is capable of illuminating some aspects of both intra-paragraph and inter-paragraph linking style. It also appears possible to make some claims concerning the development and usage of important concepts within a text. However it is not clear to what extent any 'surface' analysis of lexical features (rhetorical connectives in this case) can provide an entirely sufficient measure of text complexity.

It would be unrealistic to suppose that any measures as simple as those outlined here could provide more than a source of comparative data. In the absence of perfect machine parsing, useful information can only be expected with 'intelligent guidance' of such analysis; in an analogous fashion to the human parsing input used by some machine translation systems. Some degree of 'guided' inspection will always be required to pick up the more subtle forms of rhetoric and cohesion; although this task too might be fairly well specified.

References

Bruce, B. et al. 'A cognitive science approach to writing'. Technical Report No. 89, Centre for the Study of Reading, University of Illinois, 1978.

Clements, P. 'The effects of staging on recall from prose'. PhD Thesis, Cornell University, 1976.

van Dijk, T.A. 'Connectives in text grammar and text logic', in Dijk, T.A. and Petofi, J.S. (eds) Grammars and Descriptions, 1977.

van Dijk, T.A. 'Relevance assignment in discourse comprehension'. Discourse Processes 2, 113-126, 1979.

Grimes, J.E. The Thread of Discourse, 1976.

Halliday, M.A.K. and Hasan, R. Cohesion in English ,1976.

Kintsch, W. The Representation of Meaning in Memory ,1974.

Lefrere, P. et al. 'Expert systems in educational technology?' To be presented at APLET Annual Meeting,London, 1980.

Meyer, B.J.F. 'The organisation of prose and its effects on memory', Ph.D. Thesis, Cornell University, 1974.

Waller, R.H.W. 'Typographic design of text.' AERA Annual Meeting, Boston, 1980.

Werlich, E. A Text Grammar of English, 1976.

Appendix A

List of staging signals or connectives used in the study

Coordinators	Subordinators
[And]	After
Also	Because
Although	Before
Anyway	Except
But	For
Consequently	How
Given	If
Hence	Like
However	That
Moreover	[To]
Nevertheless	Until
Nonetheless	When
Notwithstanding	Where
Therefore	Which
Though	Whilst
Thus	Who
Unless	
Whereas	
Yet	